Data Informatics and Tools for Phase-Based Property Data

C. E. Campbell and U. R. Kattner
Material Measurement Laboratory, Materials Science and Engineering Division
A. Dima, D. Foxvog, P. J. Dessauw, and P. Savoritto
Information Technology Laboratory, Software and Systems Division

2nd World Congress on Integrated Computational Materials Engineering

Salt Lake City, Utah

July 9, 2013
Our Collaborators

NIST Information Technology Laboratory
Information Systems Group
Y.-S. Li-Baboud, B. Long

NIST Material Measurement Laboratory
Technical Program Director for Materials Genomics
J.A. Warren
Thermodynamics and Kinetics Group
B.P. Burton, R.H. Taylor, C.A. Becker, F. Tavazza, Z. Trautt
Cell Systems Science Group
T.N. Bhat, J.T. Elliott

Kent State University, Center for Materials Informatics
Laura Bartolo
Materials Genome Initiative: The Need for Data and Informatics

Goal: decrease the cost and time-to-market by 50%

Data Informatics and Tools

DATA
(Experimental and Calculated)

Tools: e.g. CMS

CALPHAD
nist.matdl.org

Reference Data

First Principles
nist.matdl.org

Atomistic Simulations
http://www.ctcms.nist.gov/potentials/

Error propagation/Uncertainty Analysis

Material Performance Criteria

Data Informatics & Tools

Materials Properties Prediction Tools

Microstructure Prediction Tools

Processing Modeling Tools

Designed New Material

Material Optimizer
Materials Are Complicated Hierarchical Systems

- Advanced materials often consist of several components (generally, $n > 5$) and multiple phases.
- The material properties are dependent on the microstructure.
- The microstructures change as a function of processing and service conditions.

Key to material design:

- What phases are present
- Composition and morphology of the phases present

Composition & phase dependent property data (CALPHAD)
Examples of Types Data

- Phase-based property data: Thermodynamic quantities, diffusion mobilities, molar volume, elastic properties, thermal conductivity, etc.
- Emphasis on binary and ternary data to predict multicomponent properties
- Data can be experimental or computational.

**1-D (Points)**
- Melting Temperatures
- Critical Temperatures (Phase Changes)
- Lattice Parameters
- Heat of Formations
- Phase fractions and compositions
- Activation energies

**2-D (Lines)**
- Composition Profiles
- Heat Capacities
- Enthalpies of mixing

**3-D**
- Crystal structures
- Micrographs/Morphologies
- 3-D Atom probe Tomography
Data

- Elements present
- Type of value (e.g. enthalpy, heat of formation, phase boundary, diffusivity, lattice parameter, bulk moduli)
  - Experimental or computational method
  - Type of measurement (direct or indirect)
- Number of phases present
- Datum value and error
  - Type (single value or series)
  - Units
  - Actual value(s) and error(s)
- For each phase present
  - Phase name
  - Composition and fraction and errors
  - Crystal structure (this input will follow the format prescribed by the CCN) or amorphous
  - Lattice parameter
- Temperature and error
- Pressure and error
- Type of Material
  - Bulk composition
  - Material purity
  - Sample preparation
  - Microstructural information
    - Single crystal
    - Polycrystalline (grain size, dislocation density)
    - Non-crystalline

Metadata

- Data manipulation details (if any, e.g. reference state corrections, analysis method to determine interdiffusion coefficient)
- Reporting format (raw data, digitized data, other)
- Reference (DOI or text; one must be present)
- Additional information
Phase-Based Property Database

Material Property Database Exist

Generally, focused on engineering/design specs or first-principle calculations results.

- Focus on phase-based properties that are needed to describe the composition, temperature, and pressure functions of a phase.
- Unary, binary and ternary data are primary focus.
- Multicomponent data are needed for validation.
Industry user just wants to know the diffusion coefficient at a given temperature for a given material.

Calphad user who wants to complete a new assessment.

Advanced expert wants to understand the diffusion mechanism.

\[ D = 2.16 \times 10^{-15} \text{ m}^2/\text{s} \]

\[ D = D_0 \exp \left( -\frac{Q}{RT} \right) \]

Data are diverse  Data are semi-structured  Need complete data sets
Architectural Strategy

Web Interface

Middleware

Workflow Engine

Schemas Documents

Ontology

Data

Future Data

Computational Resources
Informatics Approach

Ontologies
XML Schemas

Extract

Low-Level Domain Language

Transform

XSLT
Reasoners

NoSQL Databases

MongoDB, Neo4J

Materials Ontology currently being developed

Note this is a work in progress
An ontology renders shared vocabulary and taxonomy which models a domain with the definition of objects and/or concepts and their properties and relations.
Benefits from an Ontological Approach

- **Semantic Unification**
  - The unification of lexically different representations that have the same semantics
    - Example: fcc phase in steels can be referred to as fcc, austenite or γ.

- **Ontology-based Data Integration**
  - Using ontologies to unify data that share some common semantics but originate from unrelated sources
    - Example: Are property data from two experiments consistent enough to be combined?

- **Ontologies are not static and can grow with needs**
• Why use a graph database?
  – True networked database with queries, ACID, and REST interface
  – All apps can share the same representation
  – Overcomes some of the limitations of RDF
  – Flexible visualization...
Sources
- Prototype MGI Ontology
- ThermoML
- MatML
- MatSeek
- UnitsML
- ChemML

Tools
- UML (Unified Modeling Language)
- Semantic Web (RDF, OWL)

Note: This is a generalized model depicting overall structure
Data Collection: Tracer Diffusivity Test Schema

Material Genome Initiative
XML Form Editor

Data Entry
In this step, you have to fill in the form. During the process, you can view every field, until you have filled every field, you can view the XML.

Experiment
- ExperimentType
  - Choose TracerDiffusivity
- TracerDiffusivity
  - Material
    - MaterialName: Mg
  - Phase
    - Name: HCP
  - CrystalStructure
    - SpaceGroup:
      - SymbolOrNumber
  - WyckoffSequence:
    - Sequence
- Composition
  - QuantityUnit: mass fraction
  - Constituents
    - Element: Ac
    - Quantity
      - Purity
      - Purity
  - Error
    - MaterialForm
      - Choose: SingleCrystalline
- Environment
  - Environment

© 2012 NIST - MGI - XML Form Editor 0.31 | Privacy Policy | Terms of Use
Why XML?

- Internationalization
- Flexibility
- Transformability
- Interoperability
- Longevity
- Web-Enabled
- Available Resources
XML Document Storage

- **MongoDB**
  - Schema-less, cloud-friendly
  - High Performance, scalable
  - Used by CERN to enable information discovery on Compact Muon Solenoid data
- Reason: “dynamic queries, full indexes, including inner objects and embedded arrays, as well as auto-sharing”

```
{
    "_id" : ObjectId("4be97eaebed10b3e86000000"),
    "title" : "Ordered List",
    "creator_id" : ObjectId("4be97eabed10b3e860000001"),
    "memberships" : [
      ObjectId("4be97eabed10b3e8600000001"),
      ObjectId("4be97eaebed10b3e860000002")
    ]
}
```
Initial Data Set: Reference Data for Self Diffusion

Data for the Pure Elements

• Need for reference diffusion data for the pure elements

• Current process:

  Search data → Digitize Data → Plot data → Reference values

Future

File Repository/DSpace

Data

Fig. 1. Gaussian penetration profile. Error bars are the standard deviation of the mean. The solid line represents the best fit of the data to a Gaussian function.
Future Data Informatics

Ontology

XML Schema

Data Capture

JSON

Data Tools: Statistics; Machine Learning

Various Database Platforms

mongoDB

Neo4j, the graph database

Java

UNIFIED MODELING LANGUAGE

RDF

Semantic Web

User Interface
Needs for the Future Data Infrastructure

Building a data infrastructure that enables complex searches and allows for data mining and machine learning

- Need methods to facilitate data capture
  - User defined templates
  - Automated capture from instruments
  - Electronic laboratory notebooks?

- Community participation
  - Incentives from journals and funding sources